Working Paper 92-31
July 1992

División de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

MEASURING THE EFFECTS OF COVARIATES ON DURATION DATA THROUGH COMPLETELY CENSORED AND LENGTH BIASED CPS-LIKE DATA

Teresa Villagarcía*

Abstract

This study shows the problems that arise when the estimation of the differences in the duration of unemployment experienced by workers with different characteristics is calculated using cross-sectional data and methods that do not take account of the longitudinal nature of the duration data. We propose an alternative method that avoids these problems using techniques of Statistical Survival Analysis.

Key words:
Length Bias, Censored Data.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

1 Introduction

This study shows the problems that arise when the estimation of the differences in the duration of unemployment experienced by workers with different characteristics are calculated using cross sectional data and methods that do not take account of the longitudinal nature of the duration data.

In section 2 the nature of the duration data and their specific problems are briefly described. Section 3 deals with the effect of using cross sectional techniqes to analyze these data. Section 4 proposes alternative practical methods to study the effect of covariates on the duration variable and provide some simulation results. Finally, section 5 uses these methods to estimate the duration of Unemployment of Spanish workers using cross sectional data from the EPA, the Spanish equivalent to the CPS. The paper ends with a concluding section.

2 The Nature of Unemployment Duration Data.

The problem of estimating the duration of unemployment has been widely studied during the past two decades, since economists became aware of the biases that arise when the duration data are collected trough cross-sectional surveys, as is the case in most western countries.

Three important problems can bias the duration estimations if the data are collected through cross-sectional surveys: length bias, censoring and non steady state conditions.

To illustrate this, Figure 1 shows the Duration Data and the way they are collected by the cross-sectional survey. Each horizontal line represents the time spent in unemployment by one worker. The survey takes place at t_s, so the population that can be sampled is composed by the workers whose unemployment is in progress at t_s -workers 1, 4 and 5-. Workers 2 and 3 can not be detected as unemployed by the survey because they are not unemployed at the survey time.

This illustrates the length bias problem: As the unemployment spells that can be registered are only the ones that are in progress at the time of the survey, the population that can be sampled by the survey is not the true one but the one conditioned to being in progress at the time of the survey. The length bias has a geometric approach since the cut of a spell by the vertical line that represents the survey is more likely to occur in the long spells than in the short ones. The sample collected in this way is called a length biased sample, and it will have an overrepresentation of long spells and a subrepresentation of short spells. Therefore, the length bias will produce an overestimation of the measures of the duration of unemployment.

The censoring problem can also be explained by Figure 1. The data collected by the survey are not the complete spells T_1, T_4 and T_5 but the censored times C_1, C_4, and C_5. Interrupted spells are always going to be less or equal than complete spells, so the measures of unemployment are going to be underestimated due to the censoring bias. The relationship between both biases -censoring and length bias- will be discussed in section 3.

Finally, when the Labor Market conditions are changing another bias arises since the distribution of durations is not constant through time. Salant (1977) and Sider (1985) study the distributions of Unemployment and consider the effect of non-stationary conditions. Both studies deal with univariate distributions.

The statistical framework used to deal with duration data is the Survival Analysis which is broadly used in areas such as reliability, biology and economics. The Analysis of Survival Data is focused on the study of positive random variables that might be censored. The nature of the censoring mechanism and the independence or not of the duration process and the censoring mechanism is the main difficulty in this field. Extensive treatments of the subject can be found in Kalbfleisch and Prentice (1980) and Lawless (1982). Kiefer (1988) surveyed the application of these techniqes to unemployment duration.

Many important works have dealt with the univariate estimation of the duration of Unemployment, and the biases that contaminate the data. Salant (1977) suggested a parametric estimation of the complete Duration of Unemployment using data from the Bureau of Labor Statistics. The nonparametric approach, using the well known Kaplan-Meier product limit estimator (Kaplan and Meier 1958, Lawless, 1982), has been also widely used (Kaitz 1970, Baker and Trivedi, 1985). Kiefer Lundberg and Neumann (1985) compared the results of the parametric and nonparametric approach.

The next step is the introduction of explanatory variables in duration models, and this has also been extensively studied. The classical regression approach is not appropriate in this context since the dependent duration variable cannot be negative and might be censored. There are however parametric regression methods that are appropriate for censored non-normal random variables. Exponential, Weibull or Gamma regression models are appropriate in this case (Lawless, 1982).

A very attractive approach is the Proportional Hazards Model first suggested by Cox (1972). This method assumes that the hazard functions among individuals are proportional to a common baseline depending on the value of the covariates. The proportional hazard model does not make any further assumption on the baseline, thus being a semi-parametric estimation procedure. It also allows for time dependent covariates (Kalbfleisch and Prentice, 1980). As it will be discussed in section 3 the Proportional Hazards assumption will not generally be true when dealing with cross sectional data.

The application of these methods is appropriate when the data arise from a panel

survey and there are not many ties. Then the observed duration data, completed or censored, can be introduced in the likelihood function without further problems. In the next two sections we are going to study regression models when all the data are censored and there is a length bias, as in cross- sectional surveys.

3 Density of the Dependent Variable when it is always Censored and Length Biased

Let T be a random variable that measures the duration of complete unemployment spells, and assume that the distribution of T does not change with time. Let $f(t)$ and $F(t)$ be the pdf and cdf of T. Let $S_T(t)$ be the survivor function of T [1].

When the data are collected trough a cross-sectional survey that censors them at the survey time we shall define the random variable Y as the observed censored unemployment spell, with pdf $g(y)$ and survivor function $S_Y(y)$.

Figure 2 shows the duration data and the way they are collected by the cross sectional survey. The cohort of unemployed workers that begin their spells in $T_1, T_2, \ldots$ have a distribution of complete spells given by $f(t)$. We assume that the probability of entering Unemployment and the distribution of Unemployment spells are constant with time. The survey takes place ar t_s, and it will register as unemployed only those spells of the $i-th$ cohort that are longer than $t_s - T_i$. The censored spell of unemployment will take the value $t_s - T_i$.

From figure 2 we can obtain $g(y)$:

$$g(y) = \frac{\int_y^\infty f(x)dx}{\int_0^\infty S_T(x)dx} = \frac{S_T(y)}{\mu_T}$$

Obviously $g(y)$ is a censored, length biased density.

Now suppose that we are interested in studying the effects of covariates on T and, not observing T, we carry on the analysis using Y as the dependent variable in a classical regression model.

In this case we have:

$$g(y \mid x) = \frac{S_T(y \mid x)}{\mu_{T|x}} \tag{1}$$

[1]The survivor function $S(t)$ is defined as $S(t) = P(T > t) = 1 - F(t)$. It is with the hazard function or failure rate $h(t) = f(t)/S(t)$ the most used function in statistical life time data analysis. A good introduction is Lawless (1982) or on its applications to unemployment duration, Kiefer (1988).

Since $S_T(y \mid x)$ is a monotonically decreasing function, and $\mu_{T|x}$ is a constant that allows $g(y \mid x)$ to integrate one, $g(y \mid x)$ is a monotonically decreasing density having

$$g(0 \mid x) = \frac{1}{\mu_{T|x}}$$

and

$$\lim_{t \to \infty} g(t \mid x) = 0.$$

Figure 3 shows the densities of $T \mid x$ and $Y \mid x$ if $T \sim N(\beta_0 + \beta_1 x, \sigma^2)$. As the figure shows, the observable relationship between Y and x is going to have a triangular shape.

To show this we have simulated a Normal and a Weibull true relationship between T and x given by:

- Normal

$$t_i = 5 + 0.1x_i + u_i$$
$$u_i \sim N(0, 2^2)$$

- Weibull [2]

$$S(t \mid x) = \exp -(\lambda(x)t)^{\gamma}$$
$$\lambda(x) = e^{-3x}$$
$$\gamma = 3.33$$

Figures 4.a and 5.a show the simulated sample from the true density for both models and $\mu_{T|x}$. Figures 4.b and 5.b show the observable sample, censoring the true sample every month. The triangular shape appears in both figures.

Now if we estimate the effect of x on unemployment duration using the censored/triangular sample we are getting biased estimators:

$$E(y \mid x) = \int_0^{\infty} u g(u \mid x) du = \int_0^{\infty} u \frac{S_T(u \mid x)}{\mu_{t|x}} du = \frac{1}{\mu_{t|x}} \int_0^{\infty} u S_T(u \mid x) du$$

and integrating by parts,

$$E(y \mid x) = \frac{1}{2} \frac{E_T(y^2 \mid x)}{\mu_{t|x}} = \frac{1}{2} \left(\frac{Var(t \mid x)}{\mu_{t|x}} + \mu_{t|x} \right) \tag{2}$$

The term $1/2\mu_{t|x}$ represents the bias originated by the censoring, whereas the term

[2] To simulate a Weibull sample we have generated a random variable Z with a Standard Extreme Value distribution. A Weibull variable can be constructed as a location-scale model with Extreme Value noise: $W = \mu + \sigma Z$

$$\frac{1}{2}\frac{Var(t \mid x)}{\mu_{t|x}}$$

represents the length bias.

It can be demonstrated (Barlow and Proschan 1967) that when the true density is DFR -Decreasing failure rate or hazard function- $Var(t \mid x) > \mu_{t|x}$. So the length bias is going to dominate and duration measures are going to be overestimated, since $\mu_{Y|x} > \mu_{T|x}$.

If the true density is IFR -Increasing Failure Rate- $Var(t \mid x) < \mu_{t|x}$, the censoring bias is going to dominate and therefore $\mu_{Y|x} < \mu_{T|x}$. In this case duration measures are going to be underestimated. Only when the true density has constant hazard, both biases are equal and the censored length biases density will be the same then the true one. The only probability model with constant hazard is the exponential model.

If the true relationship were normal, $var(t \mid x) = \sigma^2 = k$ and $\mu_{t|x} = x\beta$ so that

$$E(y \mid x) = 1/2(\frac{k}{x\beta} + x\beta)$$

So, if β_Y is the estimated β using Y as the dependent variable we see that $\beta_Y \rightarrow 1/2\beta$ as x icreases. In Figure 5.a the lines represent the OLS estimation using the censored sample, and the authentic regression line.

The conclusion is that if we try to estimate β from cross- sectional, length biased data we are getting biased estimations of the parameters. The bias is given by equation 2.

3.1 Proportional Hazards Model

A very popular approach to studying the effect of covariates on a duration variable is Cox's Proportional Hazards Model.

A PH family is a class of models with the property that different individuals have hazard functions that are proportional to one another. This implies that the hazard function of T given x can be writen in the form

$$h(t \mid x) = h_0(t)\Psi(x\beta) \tag{3}$$

where $h_0(t)$ is a baseline hazard function, being the hazard function for an individual with $\Psi(x\beta) = 1$.

If T follows a PH model,

$$S(t \mid x) = S_0(t)^{\Psi(x\beta)}$$

$$f(t \mid x) = f_0(t)\Psi(x\beta)(S_0(t))^{\Psi(x\beta)-1}$$

Under this assumption the estimation of β can be done without any further assumption on $h_0(t)$ (Cox 1972, Kalbfleisch and Prentice 1980).

If 3 holds the density of Y is given by

$$g(y \mid x) = \frac{1}{\mu_{T|x}} S_0(y)^{\Psi(x\beta)}$$

witch, in general, is not going to be a Proportional Hazard model.

Figure 6.a shows the hazards of three censored Weibull [3] variables, with $\lambda_1 = e^{-4}$, $\lambda_2 = e^{-4.5}$ and $\lambda_3 = e^{-5}$. Figure 6.b shows the hazard ratios which are not constant and thus the censored variable is not PH.

4 Alternative Approach

Since we know the density of the censored variable, $g(y \mid x)$, the likelihood function will be:

$$L = \prod_{i=1}^{n} g(y_i \mid x_i) = \prod_{i=1}^{n} \frac{1}{E_T(t \mid x_i)} S_T(y_i \mid x_i) \tag{4}$$

This likelihood function uses the survivor function of the authentic variable. This is a great advantage since usually one knows more on the nature of the true variable then on the nature of the censored one. To obtain the parameters estimators is very simple using numerical methods such as Newton-Raphson.

In the exponential case, as $S_T(y \mid x) = exp(-y/\lambda(x))$, and $E_T(t \mid x) = \lambda(x)$, the likelihood 4 is the same that we would have obtained with the complete data.

The diagnosis of the model can be made through the generalized residuals defined by:

$$e_i = H(y_i \mid x_i) \tag{5}$$

where H is the cumulative hazard function (Cox and Snell 1968). It is easily shown that e_i are i.i.d. with standard exponential distribution.

[3] A weibull model has $h(t \mid x) = \lambda(x\beta)\gamma(\lambda(x\beta)t)^{\gamma-1}$. Clearly $\frac{h(t|x_1)}{h(t|x_2)}$ does not depend on t, thus being a PH model.

It is important to note that any definition of residuals based on $E(t \mid x\beta)$ is not of interest because these residuals are length biased in the sample.

A very common type of Unemployment Duration Data arises from a cross sectional survey were the employed workers are asked about the duration of their last unemployment spell. In this case there are two groups of data: the complete retrospective spells, and the censored, length biased spells that are in progress while the survey is taking place.

In this case, the likelihood function is,

$$L = \prod_{i \epsilon \mathcal{C}} g(y_i \mid x_i) \prod_{i \epsilon \mathcal{R}} f(t_i \mid x_i)$$

$$= \prod_{i \epsilon \mathcal{C}} \frac{1}{\mu_{T|x_i}} S_T(y_i \mid x_i) \prod_{i \epsilon \mathcal{R}} f(t_i \mid x_i) \tag{6}$$

where $\mathcal{C}$ holds for the Censored set and $\mathcal{R}$ for the Retrospective Complete set. In this case it is possible to fit a parametric regression model for the Restrospective complete set and use the parameter estimations as initial values for equation 6 [4].

Finally, another very usual type of Unemployment Duration Data arises from cross sectional surveys when the duration data are grouped. In this case, some authors have studied the effect of covariates on Unemployment Duration by defining a Bernouilli variable ,Z, given by [5] :

$$z_i = \begin{cases} 1 & y_i > Y \\ 0 & y_i < Y \end{cases} \tag{7}$$

so that $z_i = 1$ represents a long spell of unemplyment, and $z_i = 0$ a short one. It is important to note that a Logit Model is not appropriate for these data as Figure 3 shows: While the long spells of Unemployment are well defined, the short ones are a mixture of eventually short spells, and eventually long spells.

A quasi-logit model can still be fitted if the probabilities of z_i are defined as follows:

[4] An example of this type of data is the Spanish ECVT survey. This survey was carried in automn 1985, and has retrospective and censored, length biased data. The effect of Unemployment Insurance on Unemployment Duration has been studied by Alba-Ramírez and Freeman (1990) using Proportional Hazards Models and a Weibull regression model. They do not take account for the length bias, thus using

$$L = \prod_{i \epsilon \mathcal{C}} S_T(y_i \mid x_i) \prod_{i \epsilon \mathcal{R}} f(t_i \mid x_i)$$

instead of equation (6)

[5] Folmer and Van Dijk (1988) defined a Bernouilli variable and applied a Logit model to a censored cross-sectional sample.

$$P(z_i = 1 \mid x_i) = \int_Y^{\infty} g(u \mid x_i)du = \frac{1}{\mu_{T|x_i}} \int_Y^{\infty} S_T(u \mid x_i du$$

$$P(z_i = 0 \mid x_i) = \int_0^{Y} g(u \mid x_i)du = \frac{1}{\mu_{T|x_i}} \int_0^{Y} S_T(u \mid x_i du$$

The Likelihood function is,

$$L = \prod_{i=1}^{n} P(z_i = 1 \mid x_i)^{z_i} P(z_i = 0 \mid x_i)^{1-z_i} \tag{8}$$

It is obvious that grouped data can be treated as a generalization of binary data.

5 Simulations

In these section we present the results of some simulations to ilustrate the problems that arise when the data come from cross sectional surveys and to check the proposed method and the residuals.

We have simulated 1000 censored Weibull samples equivalent to the one showed in Figure 5.b. The histogram of $\hat{\beta}$ ($\beta = -3$) is showed in Figure 7.

Figure 8 shows the relationship between the $\hat{\beta}$ estimated from a complete sample, and the $\hat{\beta}$ estimated through a censored sample, using always the same sample in the censoring process. The line corresponds to the 45 degrees line.

Figures 9.a and 9.b show the plot of generalized residuals of the censored sample against the independent variable, and $\mu_{T|x_i}$. Figure 9.c presents the histogram of the residuals, and the standard exponential density.

We have introduced an outlier in the sample to check how the residuals can detect it. Without the outlier, the estimations of β and γ are $\hat{\beta} = -3.02$, $\hat{\gamma} = 3.05$. With the outlier ($x = 0.5862$, $y = 15$ instead of $x = 0.5862$, $y = 1$), $\hat{\beta} = -2.92$, $\hat{\gamma} = 2.15$. Figures 10.a and 10.b show the plot of the data and the plot of the residuals with the outlier clearly marked.

6 Some Real Data: Duration of Unemployment in Spain

The data that we are using come from the Encuesta de Condiciones de Vida y Trabajo en España (ECVTE). This survey was carried out in automn 1985 and has circa 61000

observations. The survey has retrospective and censored length biased data.

Table 1 gives some location measures for the Unemployment Duration data in the sample. The existence of length bias is clear given the overestimation of the location measures from the censored sample.

TABLE 1
Location Measures for the Duration of
Unemployment in months.

	MEAN	MEDIAN	n
Censored	23.70	14	2131
Complete	15.44	8	2345

Finally, Table 2 gives the estimated models for the complete data, all data using a Weibull regression model without the correction term, and the model proposed adjusted to 292 randomly chosen censored data. The two independent variables are SEX (1 if MAN) and FIRED (1 if Fired from the last job). The differences in estimations are clear.

TABLE 2
Estimated Models.

	SEX	FIRED
Complete	-.27	.29
n=2259	(0.06)	(0.06)
All	-1.04	.33
n=2017+2259	(0.06)	(0.07)
Censored	-.26	.27
n=292	(.08)	(.09)

7 Conclusions

This study shows the biases that arise when estimating the effect of covariates on completely censored length biased variables.

These data are usual in unemployment surveys, but also in reliability where the life of a component can be estimated using data from in-use components.

The usual way to estimate this, through completely censored data or mixed complete and retrospective data, introduces a bias that can be avoided using the alternative maximum likelihood method given in section 4. With this approach it is possible to model the authentic variable using completely censored data, and avoiding the biases that arise through the use of censored data.

We have also showed that the popular Propotional Hazards Model is not appropriate in this case because the density of the censored variable is not in general going to be Proportional Hazard.

References

[1] Alba-Ramlrez, A. and Freeman, R.B.(1990) Jobfinding and Wages when Longrun Unemployment is Really Long: The case of Spain *Working Paper Series. NBER*

[2] Baker G.M. y Trivedi P.K. (1985) Estimation of Unemployment Duration from Grouped Data: A comparative Study. *Journal of Labor Economics.* **Vol 3** N'2.

[3] Barlow y Proschan F. (1967) *Mathematical Theory of Reliabi- lity.* Wiley.

[4] Cox D.R. (1972) Regression Models and Life Tables (with Discusion).*J. R. Stat. Soc. B.* **34** 184-220.

[5] Folmer H. and Van Dijk J. (1988) Differences in Unemployment duration: a regional or a personal problem?. *Applied Economics* **20.** 1233-1251.

[6] Kaitz H.B.(1970) Analyzing the Length of Spells of Unemploy- ment. *Monthly Labour Review.* **Vol 93**. Nov 1970.

[7] Kalbfleisch J.D. Y Prentice R.L. *The Statistical Analysis of Failure Time Data.* Wiley.

[8] Kaplan E.L. Meyer P.(1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statis- tical Asotiation.* **Vol 53**. Junio 1958.

[9] Kiefer N.M. Economic Duration Data and Hazard Functions. *Journal of Economic Literature.* **Vol 26** pp 646-679.

[10] Kiefer N.M. Lundberg S.J. Y Neumann G.R.(1985) How long is a spell of Unemployment?. Illusions and Biasses in the use of CPS data. *Journal of Business and Economic Statistics.* **Vol 3** N'2.

[11] Lawless J.F. (1982) *Statistical Models and Methods for Lifetime Data.* Wiley.

[12] Salant S.W. (1977) Search Theory and Duration Data: A Theory of Sorts. *The Quarterly Journal of Economy.* Feb 1977.

[13] Sider H. (1985) Unemployment Duration and Incidence: 1968-1982.*American Economic Review.* Junio 1985.

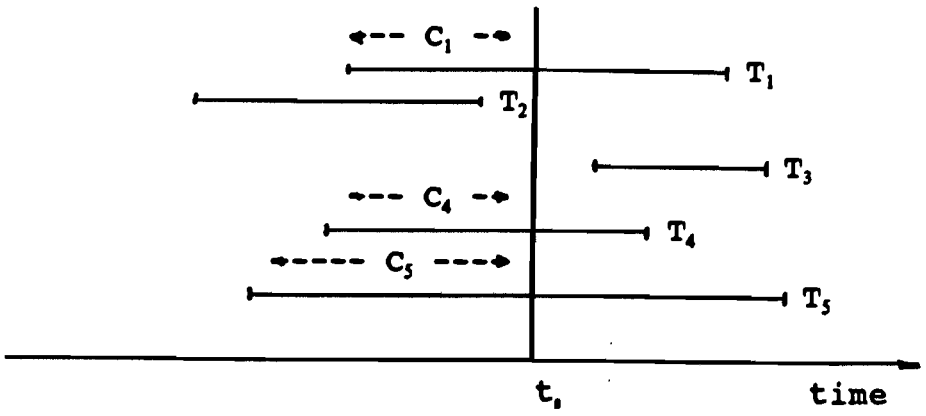

FIGURE 1

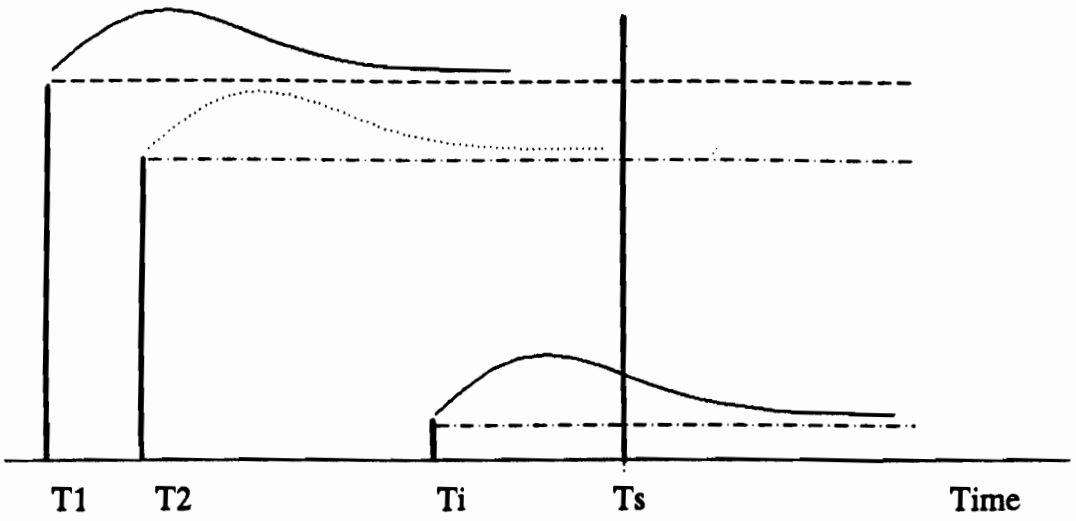

FIGURE 2

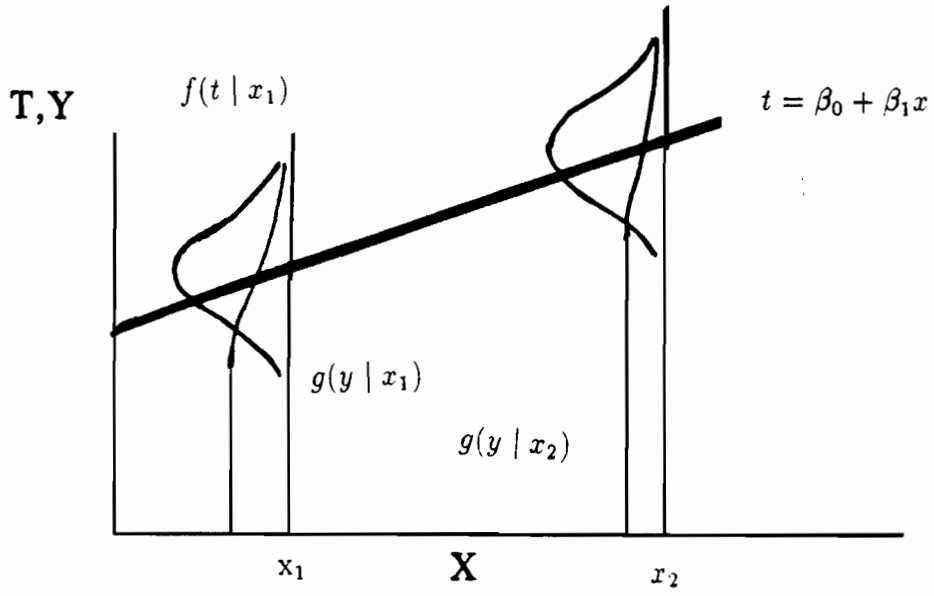

FIGURE 3

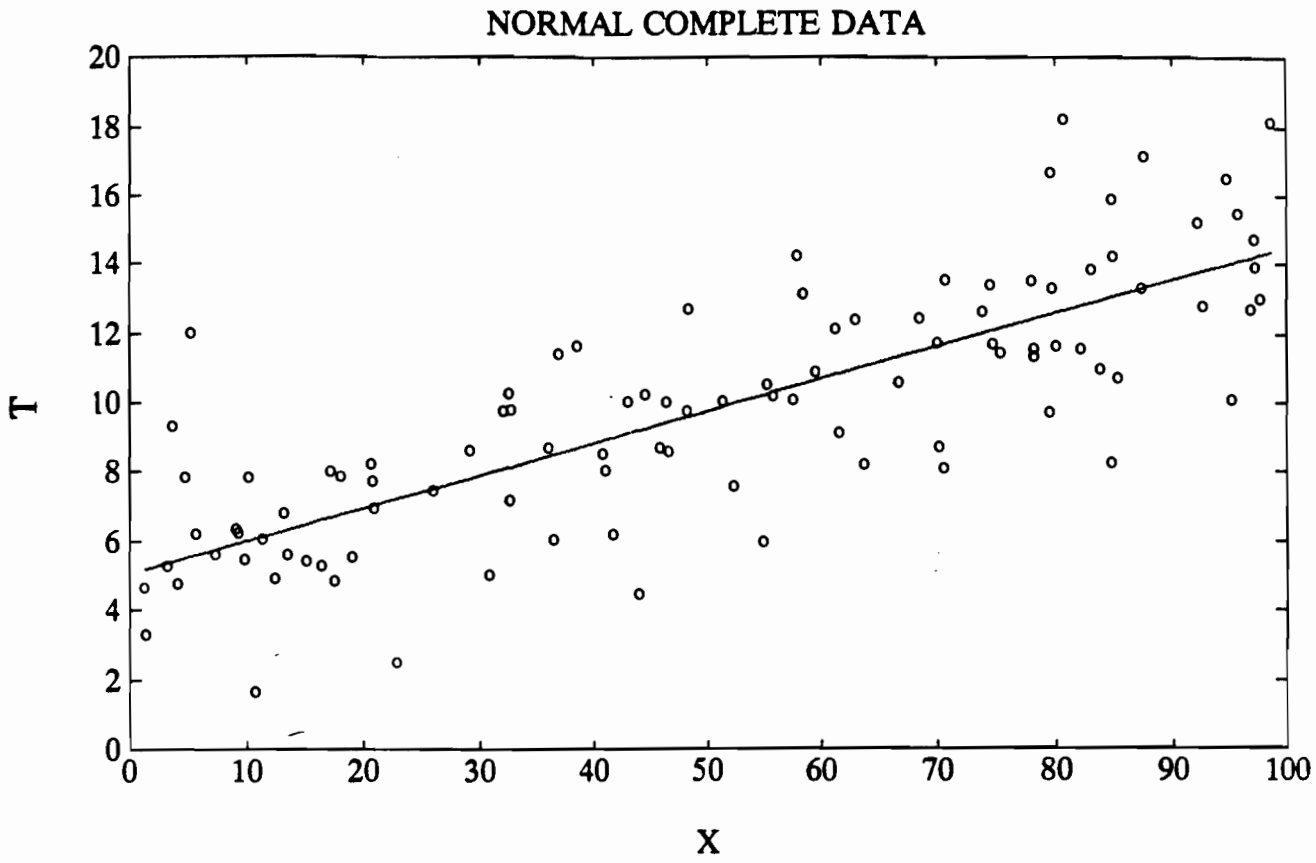
NORMAL COMPLETE DATA
T
X
0
2
4
6
8
10
12
14
16
18
20
0
10
20
30
40
50
60
70
80
90
100

FIGURE 4.a

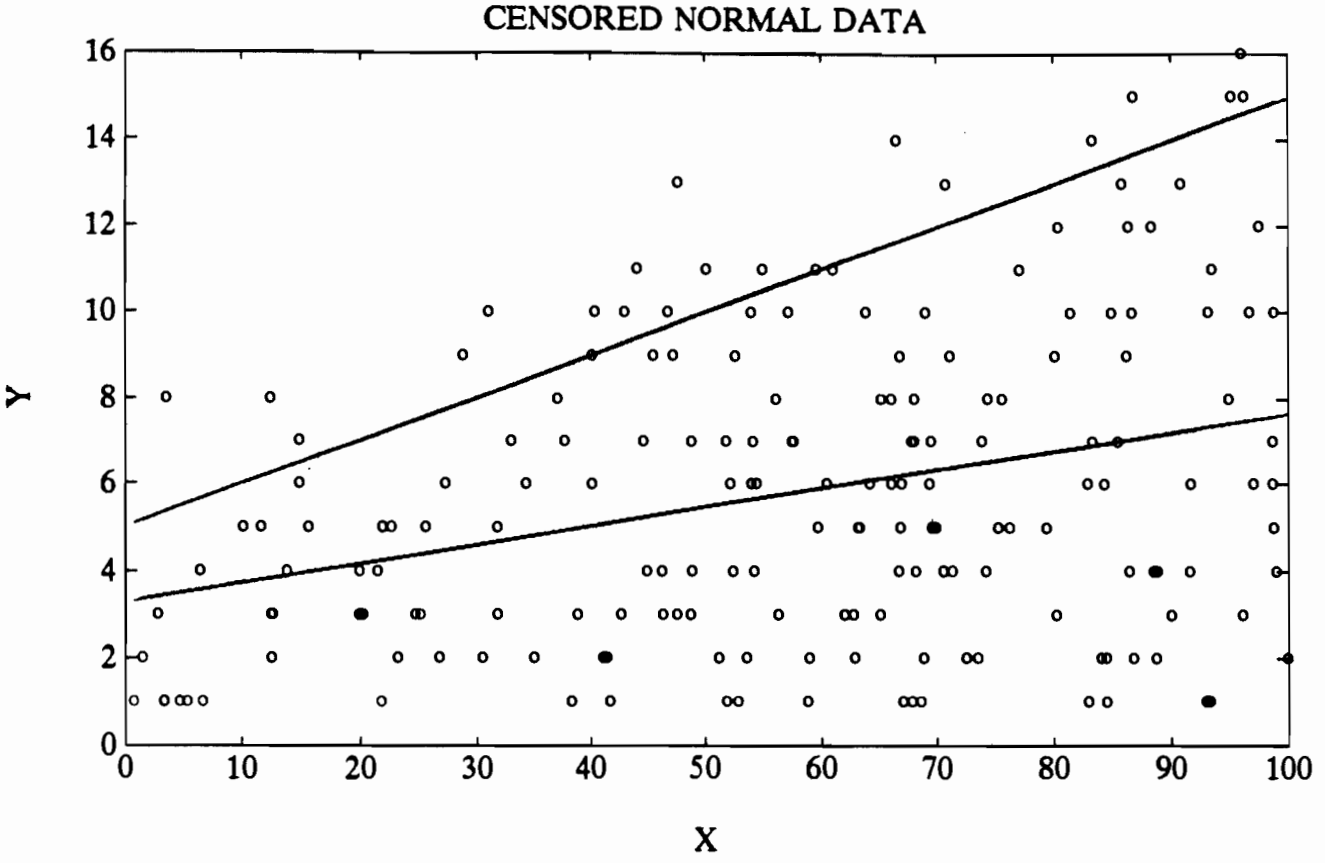
CENSORED NORMAL DATA
Y
X
0
2
4
6
8
10
12
14
16
0
10
20
30
40
50
60
70
80
90
100

FIGURE 4.b

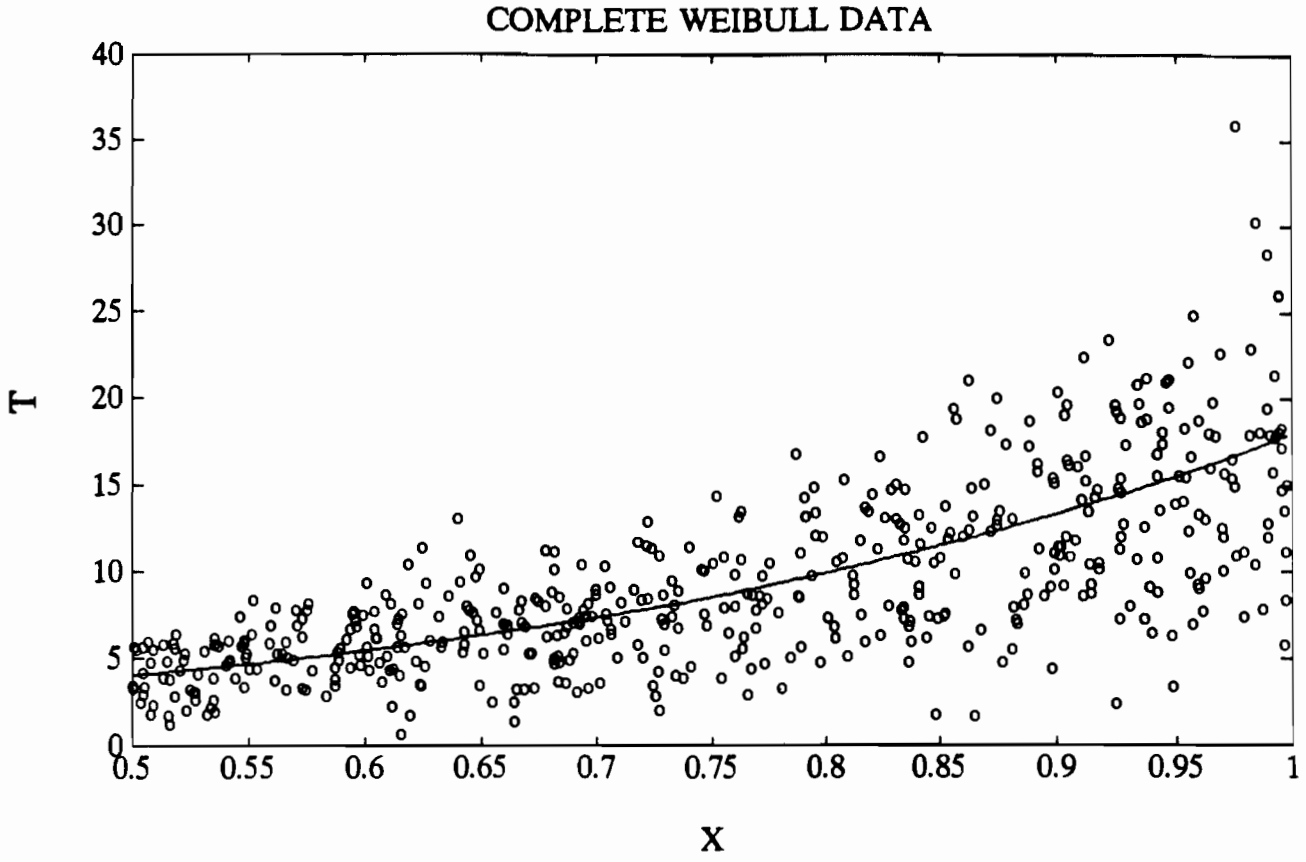
COMPLETE WEIBULL DATA
T
X
40
35
30
25
20
15
10
5
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

FIGURE 5.a

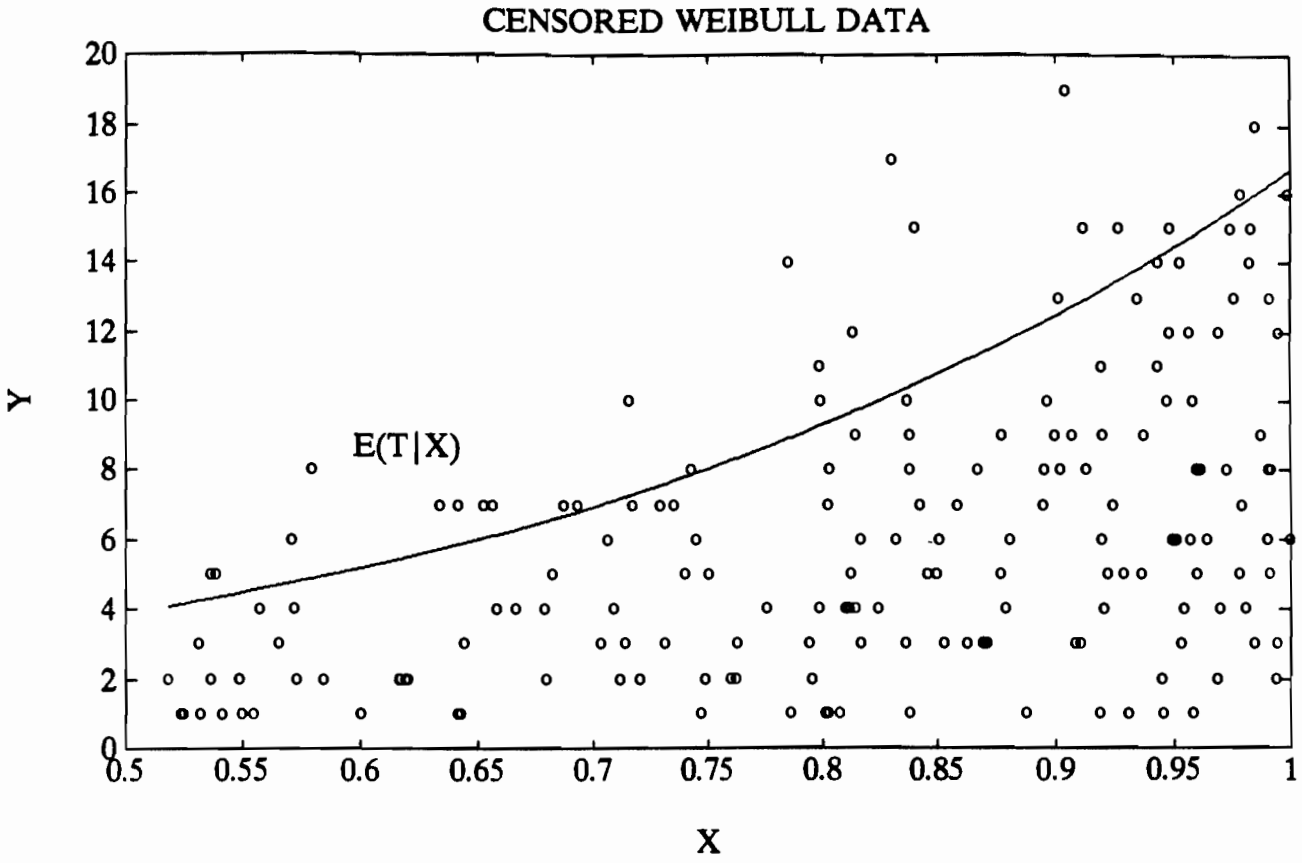
CENSORED WEIBULL DATA
Y
X
E(T|X)
0
2
4
6
8
10
12
14
16
18
20
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

FIGURE 5.b

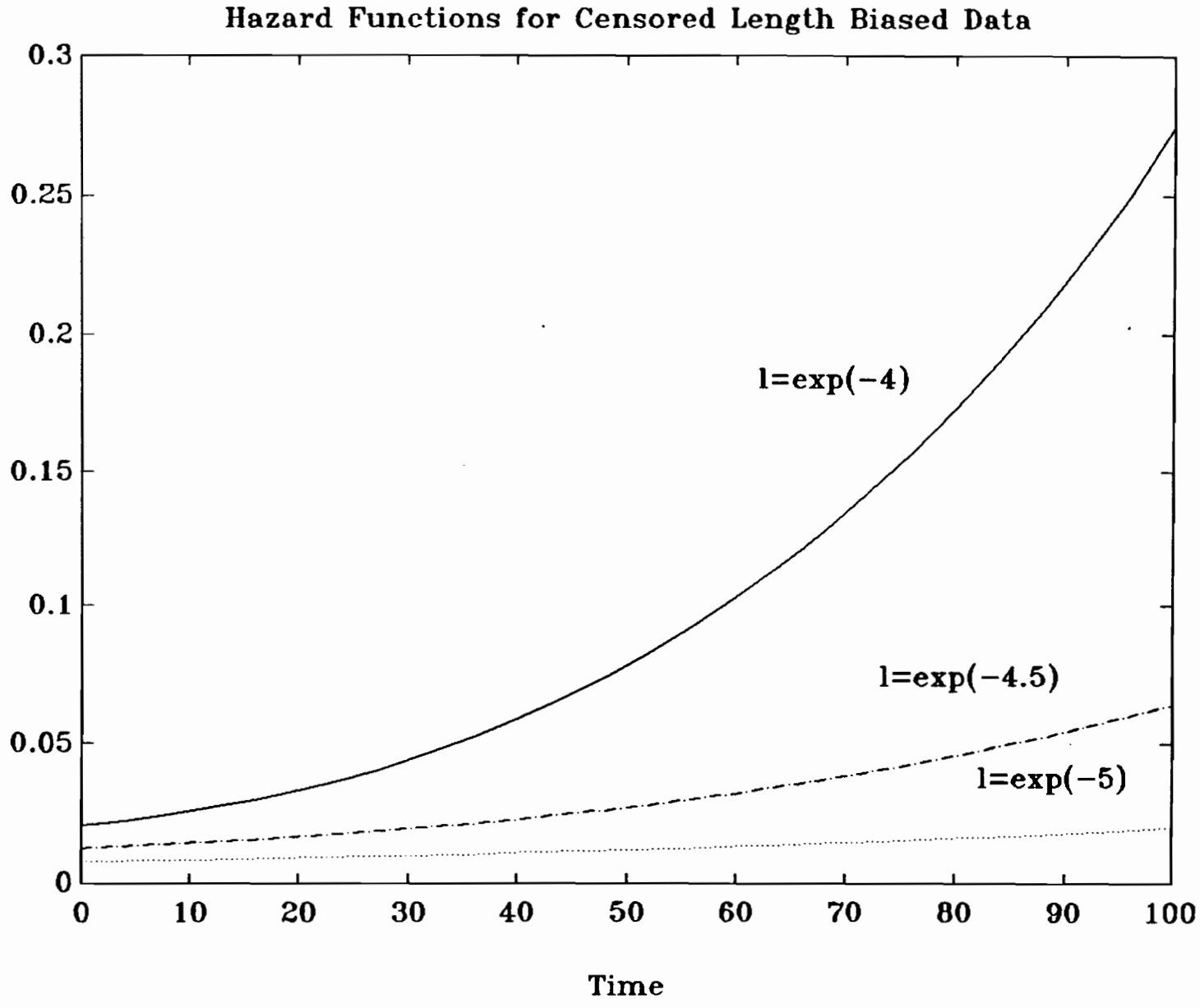
Hazard Functions for Censored Length Biased Data
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100
Time
l=exp(−4)
l=exp(−4.5)
l=exp(−5)

FIGURE 6.a

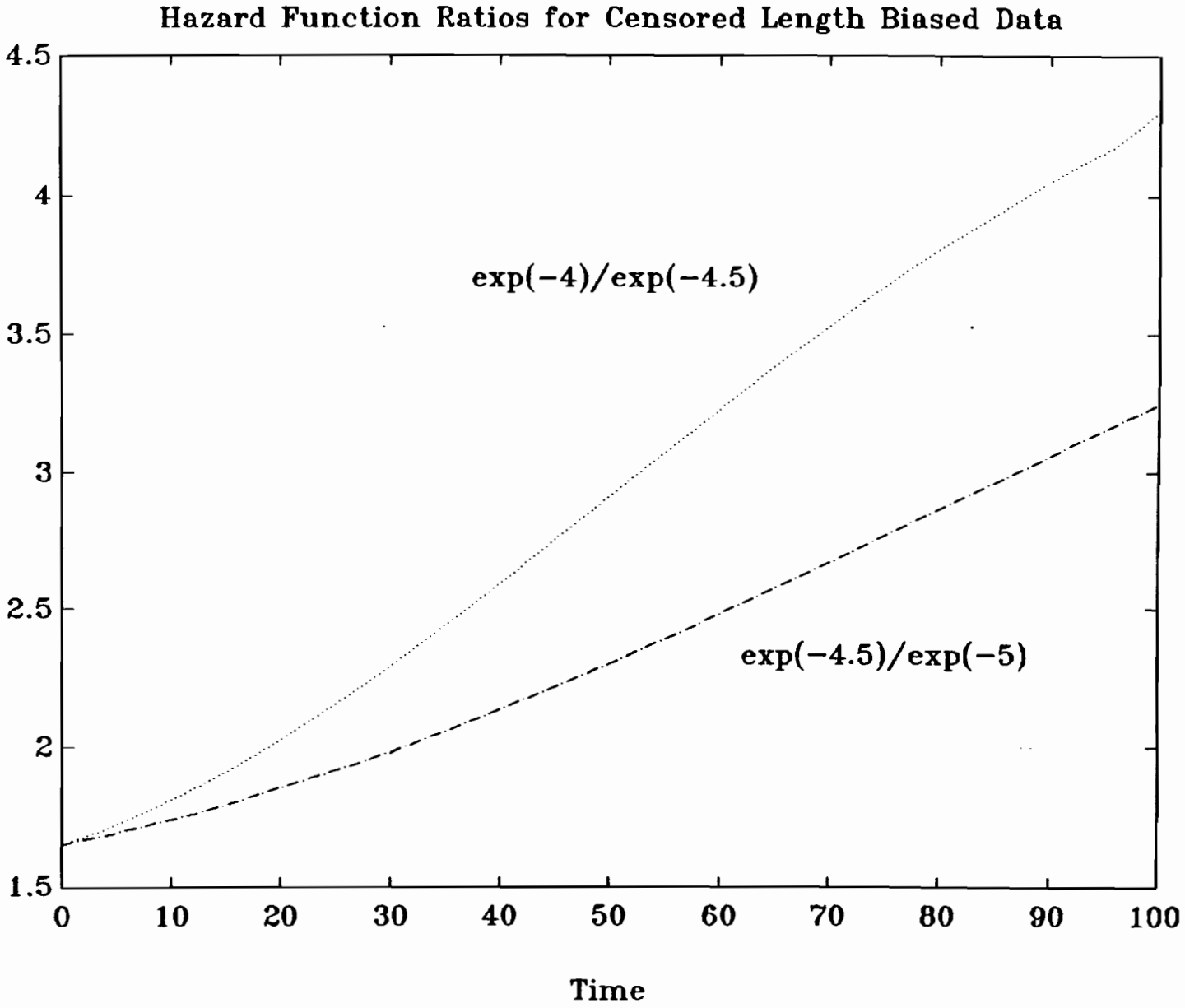
Hazard Function Ratios for Censored Length Biased Data
4.5
4
3.5
3
2.5
2
1.5
0
10
20
30
40
50
60
70
80
90
100
Time
exp(−4)/exp(−4.5)
exp(−4.5)/exp(−5)

FIGURE 6.b

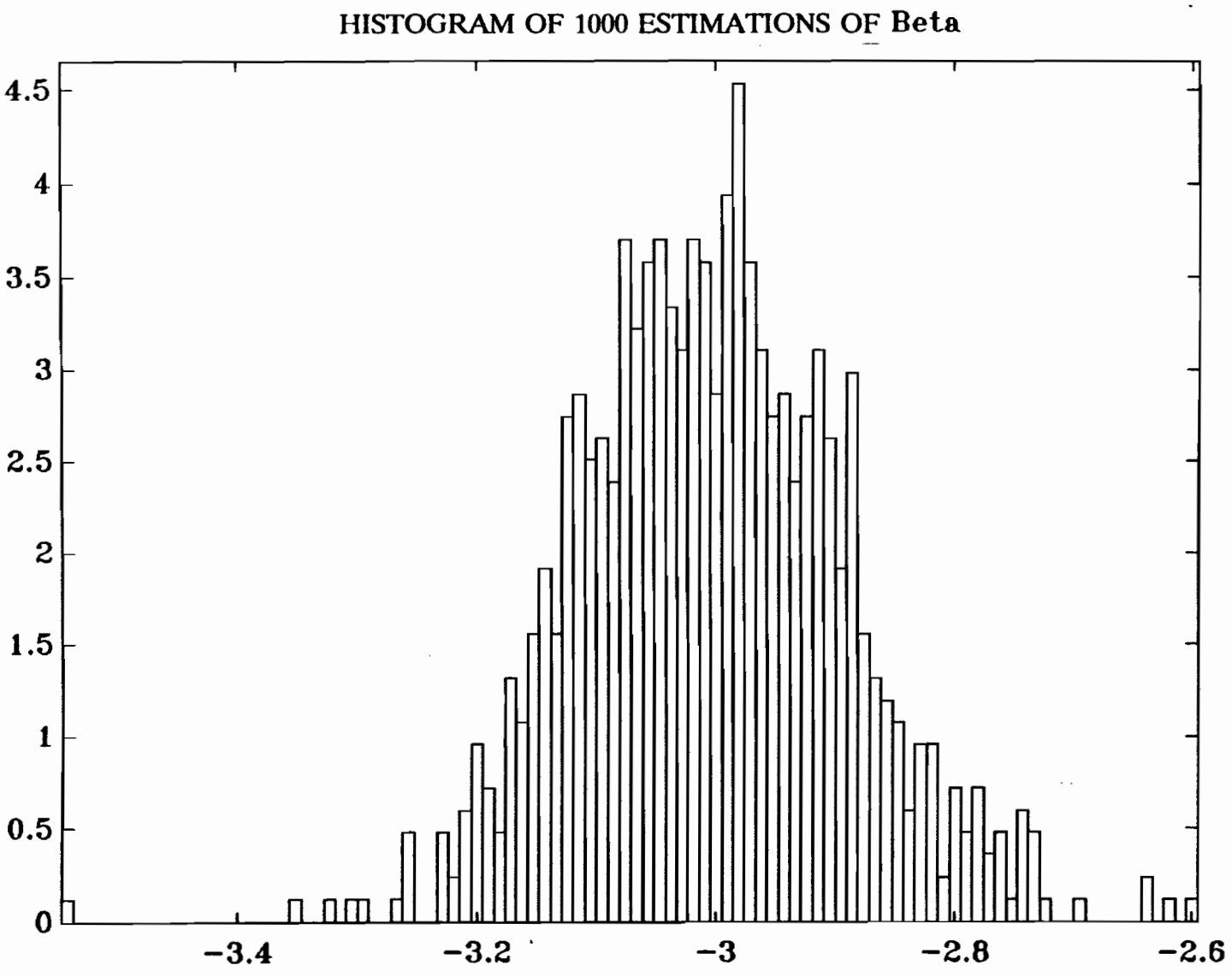
HISTOGRAM OF 1000 ESTIMATIONS OF Beta
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
−3.4
−3.2
−3
−2.8
−2.6

FIGURE 7

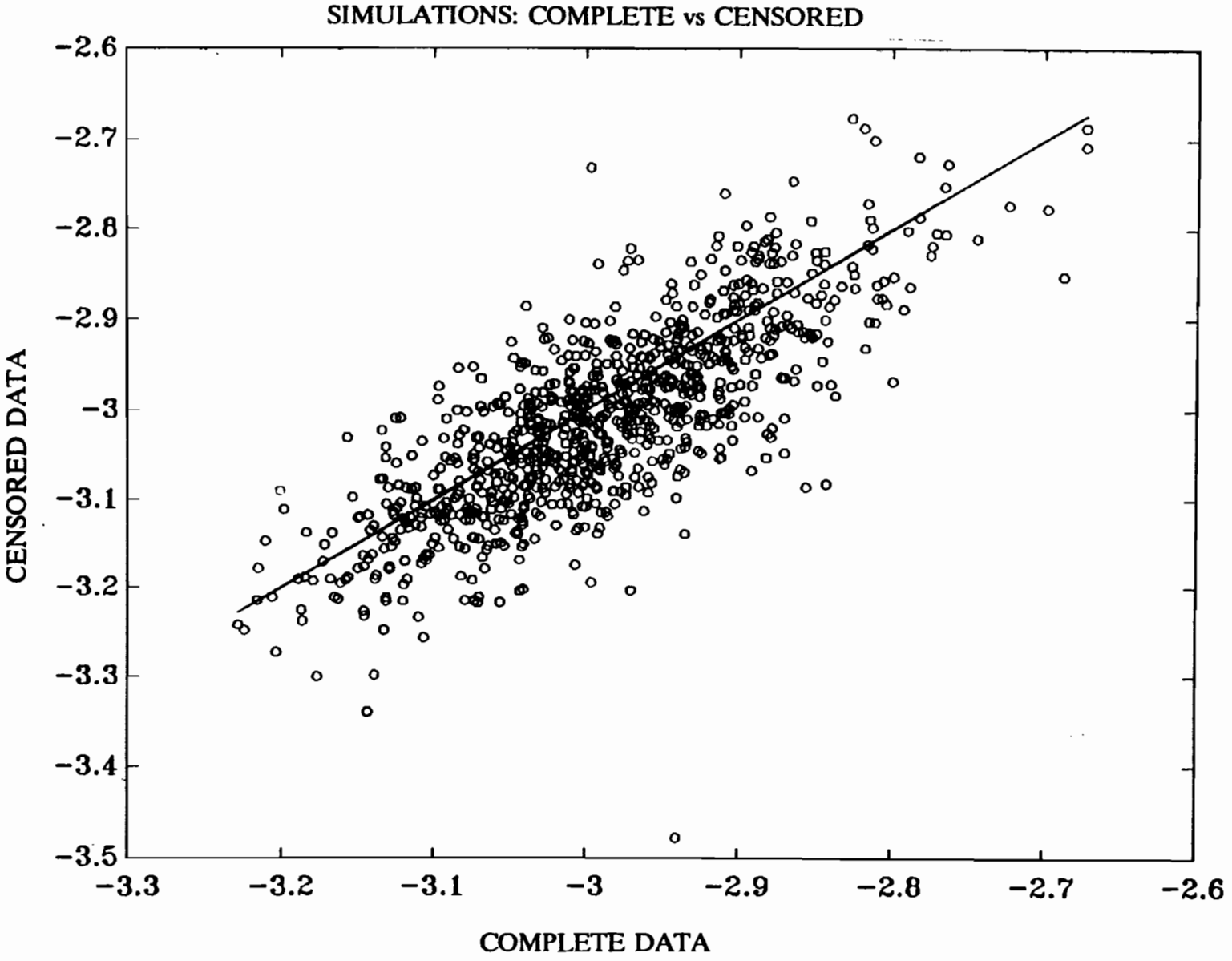
SIMULATIONS: COMPLETE vs CENSORED
CENSORED DATA
COMPLETE DATA
−2.6
−2.7
−2.8
−2.9
−3
−3.1
−3.2
−3.3
−3.4
−3.5
−3.3
−3.2
−3.1
−3
−2.9
−2.8
−2.7
−2.6

FIGURE 8

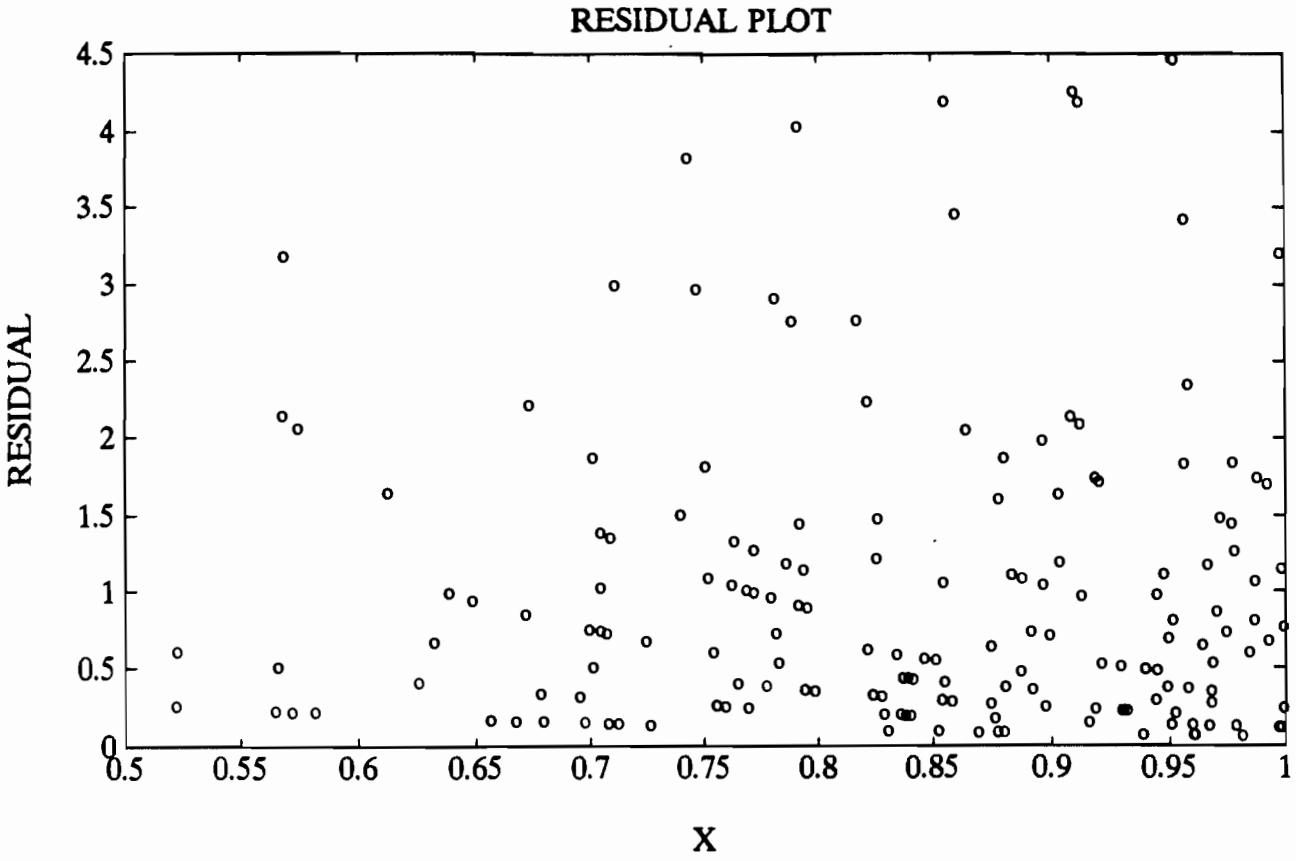
RESIDUAL PLOT
RESIDUAL
X
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

FIGURE 9.a

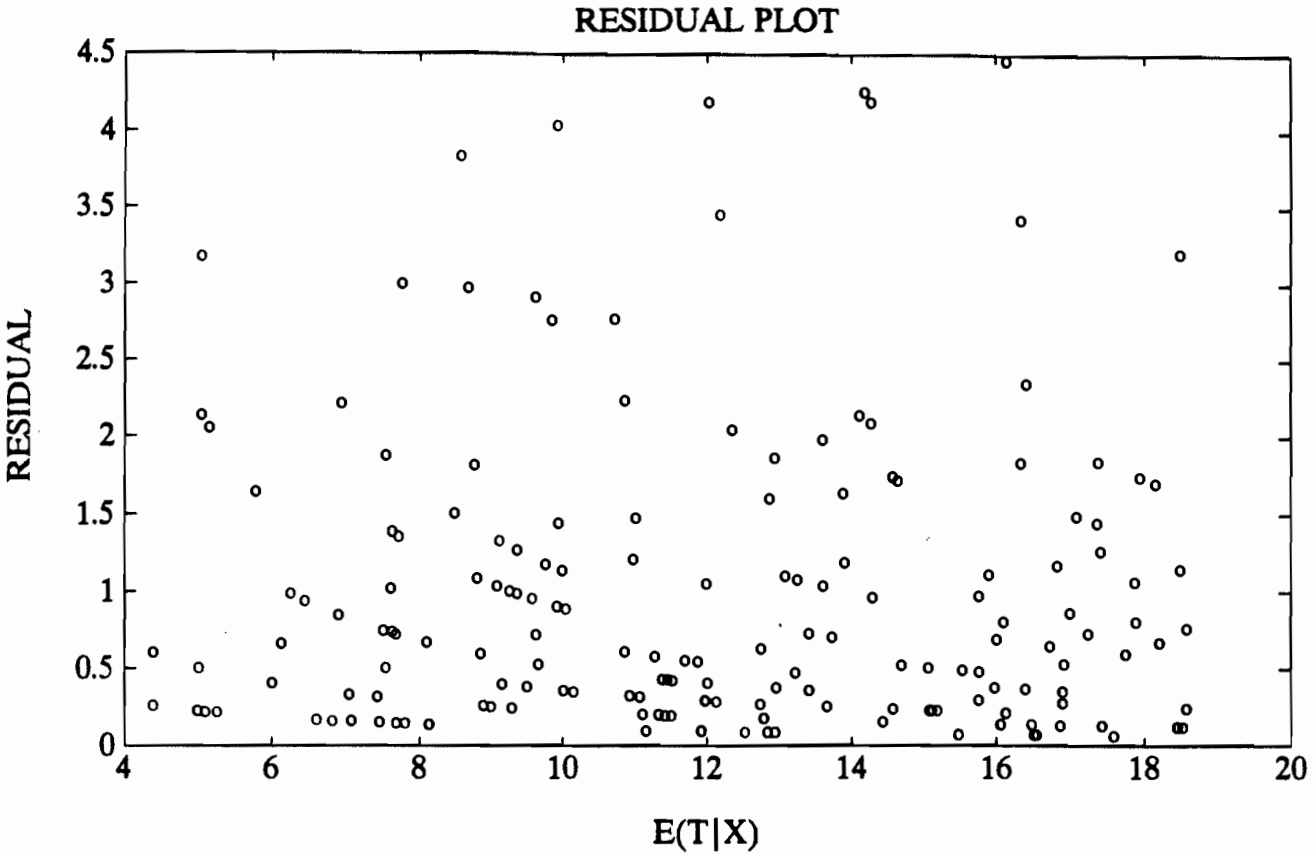
RESIDUAL PLOT
RESIDUAL
E(T|X)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
4
6
8
10
12
14
16
18
20

FIGURE 9.b

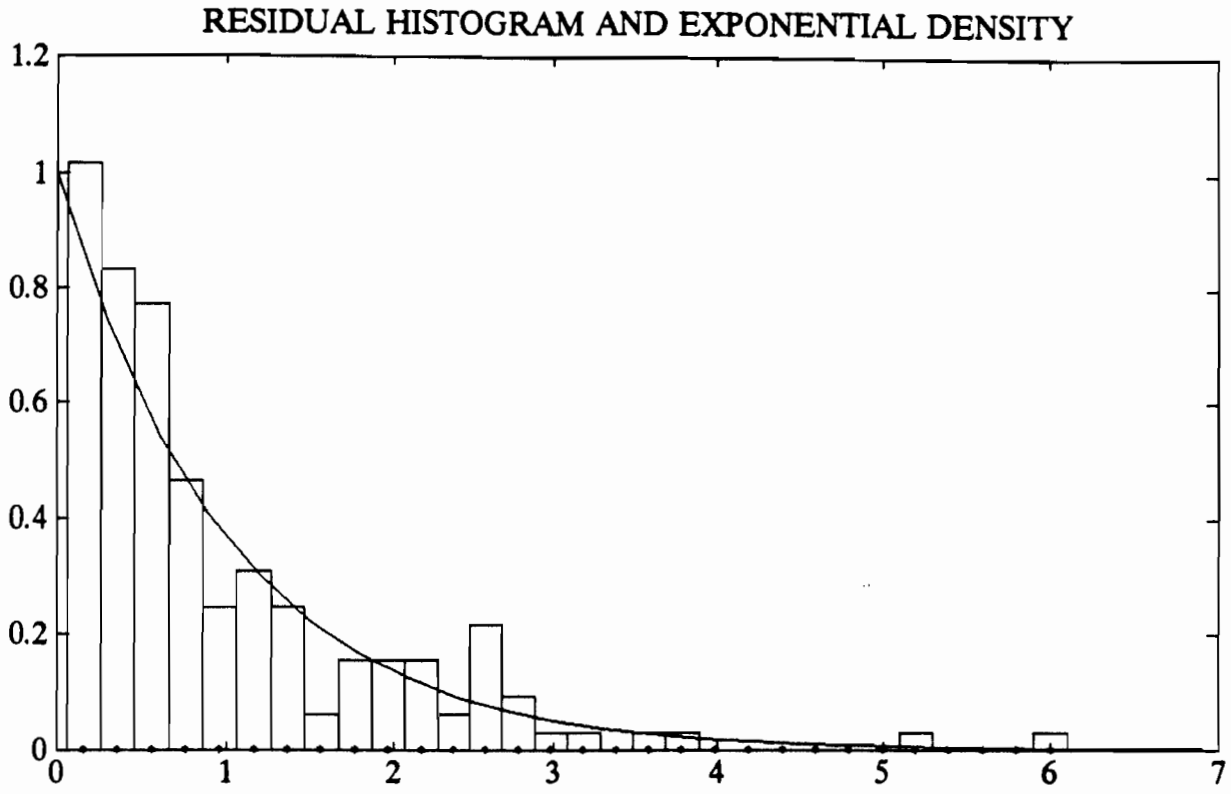
RESIDUAL HISTOGRAM AND EXPONENTIAL DENSITY
1.2
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7

FIGURE 9.c

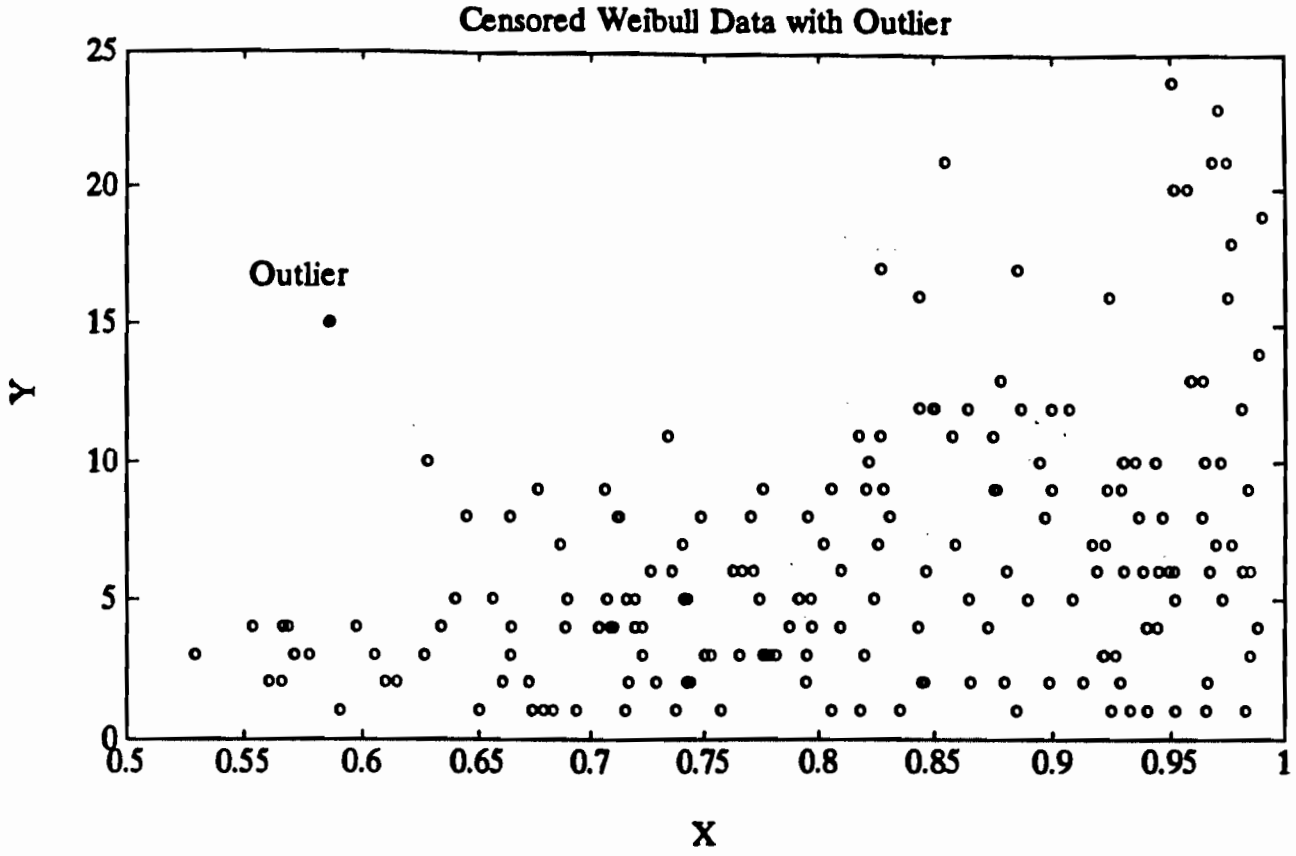
Censored Weibull Data with Outlier
Outlier
Y
X
25
20
15
10
5
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

FIGURE 10.a

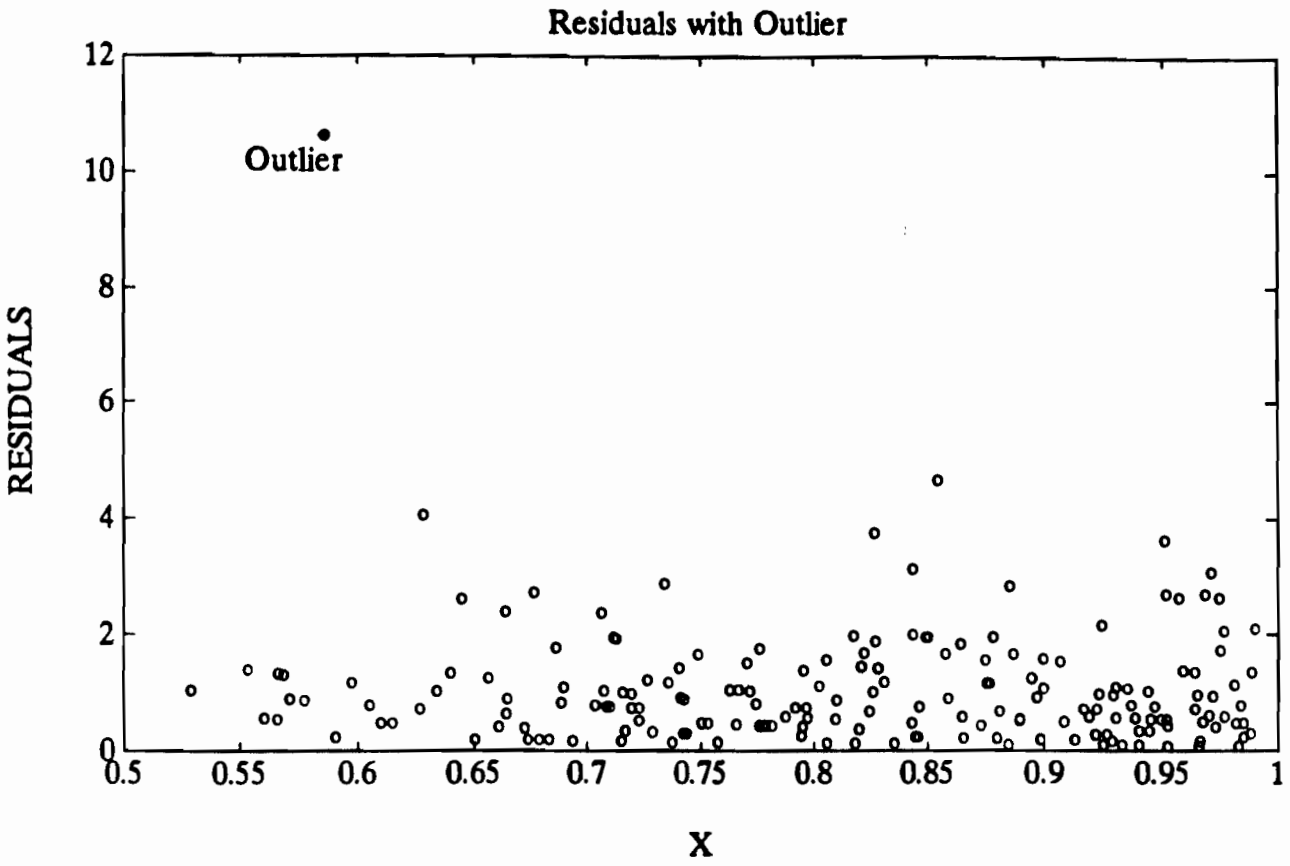
Residuals with Outlier
Outlier
RESIDUALS
X
12
10
8
6
4
2
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

FIGURE 10.b